A Data User's Guide

A Guide to Understand and Use Data

Acknowledgements

I would like to acknowledge the contributions of those who provided ideas, time, red ink, and in one instance, a donut, to make this a better guide.

James Aydelotte
Bureau of Health Policy and Vital Statistics
Division of Health
Idaho Department of Health and Welfare

October 2004

Table of Contents

Introduction

Employees at many levels of government are facing an increasing need to use data in the decision-making process. The pressure to use data manifests itself in many ways. Federal grant applications, program evaluations, program justifications, internal decision making, policy analysis and policy making, community partnering, and legislative requests are some of the ways that data are used. As the need to use data has increased, more and more employees are required to evaluate data, communicate with data analysts, analyze data, communicate data, and make decisions based on data. The difficulty lies in the fact that many people do not have the knowledge or skills to effectively take on these tasks. The result can be inefficient, ineffective, or simply incorrect interpretation and use of data. This in turn can lead to poor decision making.

The intent of this document is to serve as a resource that employees of all levels can use to gain basic knowledge and skills regarding the evaluation, analysis, and communication of data with the intent that program managers and employees can more effectively use data to support their programs.

This guide uses as its structure a generic framework for asking questions, getting answers, and reporting the answers. This framework was developed after reviewing literature on policy analysis, program evaluation, geographic information systems, and research.

The topics covered in this guide are:

- Begin with a question.
- Find the data.
- Understand the data.
- Create an analytic plan.
- Obtain the results.
- Report the data.

Included at the end are some additional resources found on the internet that may be useful to you in your search to find answers through data. This document is intended to be dynamic in its content. Additional topics that readers will find informative will be added as needed; therefore, your feedback is invited.

Why Start With A Question?

Data analysis should begin with a question because it should be systematic. Three good reasons that analysis should be systematic are:

1. It should be grounded in knowledge.
2. It should be transparent.
3. It should be focused.

Data analysis should be grounded in knowledge. A solid understanding of the issue is important in order to know what questions should be asked. A poor grasp of the issue could result in asking a question that has already been answered, thereby wasting time and money. Worse yet, it could result in framing an issue incorrectly and result in a misinterpretation of the results. A good understanding will lead to better questions. Furthermore, understanding the issue provides the context for using the results of the data analysis. Data are not meant to supplant other kinds of knowledge but are meant to complement and provide a check on it. For example, suppose that the results of an analysis are not consistent with generally held belief about the topic. The results can be used to re-evaluate common wisdom about the subject and the common wisdom can be used as a check on the results. In both cases, a grounding in the subject is vital. Understanding the issue can include the history of the issue, current status and knowledge of the issue, and the political nature of the issue.

Data analysis should be transparent. Data analysis is often pioneering work and can break new ground in a number of ways. Transparency means that the assumptions and decisions made about the analytic process and the data are well documented. Good documentation allows:

- Others to understand any conclusions that are made.
- The work to be replicated.
- The assumptions and decisions to be referenced in the future.

Data analysis should be focused. If the analysis is not focused, it will be difficult to determine when the process is done. A good question will focus the analysis. A question like, “What can you tell me about births in Idaho?” leaves a lot of room for interpretation, and the answer may not be the one originally sought. However, a specific question like, “How many births were there in Idaho in 2003?” will receive a specific answer.

How To Find the Question

Sometimes, the question will be obvious. Someone may ask you, it could be found in grant objectives or a grant guidance, or it may come from some other source. Often, however, a question needs to be developed. Usually, the general topic is already known. If so, then the question needs to be asked, "What do I need to know about _____?" The answer can be general or specific. The answer may include one or more of the following:

- The magnitude (how big or small).
- The trend(s) (what is going on over time).
- The demographics (examining by age, sex, income, education, etc.).
- The geographic distribution (looking at county, region, district or other geographic subdivisions).
- Relationships between variables (are there variables that are associated, does one impact another?).

DEFINITION: variable - a data item of interest. It is called a variable because its value can vary by record or observation. Variables are sometimes referred to as "data elements."

Knowledge of the issue becomes very useful at this point in order to pose a useful question. Program managers and others already working in a particular field may already have extensive and relevant knowledge because of education and work experience. It would also be useful to do a literature review to see how others have researched and reported on the particular issue and to find out recent developments in the field. Other ways to find useful information are to talk with colleagues, counterparts in other states, review publications (including the section citing other literature), and review information on the internet.

For example, let us say that you are interested in Medicaid expenditures (general topic). What I want to know specifically is how Medicaid expenditures for the delivery of babies has changed over time. That is a question that can be researched. It is quite possible that as you uncover more information, your question will become more specific, and you may want to look at other elements of Medicaid expenditures.

As you develop your question, identify how its answers will align with the vision, mission, goals, and objectives of the Department, Division, or Bureau in which you work. Doing so will help you demonstrate the need and relevance of your work if it needs to be justified.

A person's knowledge about an issue can encompass knowledge about a data source that will provide the answer(s) to the question(s). If not, the following suggestions will be useful.

First, continue to research the issue. The state library is a good resource. Literature searches and searches on the internet can yield valuable results. The likelihood that some research on the issue has already been done is high. If so, you may be able to use that data source or a similar one. Contact those in the community of professionals that deal with the issue. They may know about useful resources; there are often independent research projects going on that may inform your research or even answer your question(s) specifically. Contact other governmental agencies whose data may apply to your research. Agencies often have a wealth of administrative data and sometimes collect population-based data. For example, the Idaho Department of Education conducts the Youth Risk Behavior Survey – a survey of health related behaviors of high school students – every other year. Web sites of governmental agencies and not-for-profit organizations will often put publications on line or provide descriptions of the publications and how to order them. This kind of research is called "secondary research."

If you still haven't found the data you need after going through these steps, you may need to consider conducting original or "primary" research. Original research can be expensive and tradeoffs between budgets and the extent of the information collected always have to be made. The process of collecting data through primary research is beyond the scope of this document. Collecting statistically valid and scientifically sound data a rigorous process. Determining that process involves additional skills and knowledge. Contact an analyst or a researcher to obtain more information on this process.

Assessing the data requires a couple of steps. The first is to evaluate the who, what, when, where, how, and why of the data collection. The answers to these questions will give you information about the origin of the data. The second is to look at the data itself. A few simple tools can tell you about the condition of the data. These steps will be most helpful to you if you are analyzing data yourself; if you are working with a data analyst who holds the data, the analyst probably has this knowledge.

The Origin of the Data

The answers to the following questions, posed in *Basic Methods of Policy Analysis and Planning*, will give you a reasonable basis for assessing the origin of the data.

- What data were collected?
  Get a list of variables and definitions from the persons responsible for the data. This is often called the codebook or data dictionary. This list will let you know if all the variables you need to answer your question(s) are available. For example, if a variable indicating the race of each person represented in the data is critical to finding your answer(s), check to see if a race variable is present in the data set.

- Where were the data collected?
  This will inform what kinds of statements you can make about the data. For example, if the data were survey data collected to determine the percentage of the Elmore county population that has diabetes (this is called the "prevalence"), then you cannot make any claims about Ada county or the rest of the state.

- How were the data collected?
  If possible, get copies of the forms or other tools used to collect the data. These can provide a great deal of information as to the specific terms used and how questions were asked. Understanding how the data were collected will also impact how you can report the results of analysis of the data. For example, suppose that you had the results of the question, "Have you ever been told by a doctor, nurse, or other health professional that you had asthma?" You could not report that X% have ever been told by a doctor that they had asthma because the question includes nurses and other health professionals as well.

- Why were the data collected?
  The purpose for which data were collected will impact the answers to all the other questions. If the purpose was substantially different than your own reasons for collecting data, you may find that its form or substance are not compatible with your research.

- When were the data collected?
  All things being equal, the more recent the data the better –unless of course your research is about finding answers from previous years. However, quality data that are somewhat dated will better serve you than recent but poor quality data. Keep in mind that the older your data are, the more likely conditions have changed enough to make conclusions drawn from the data invalid.

- Who collected the data?
  Finding out who collected the data is important for two reasons. The first is a check on the quality of the data. If the organization that collected the data has a particular agenda or a questionable research history, you should pay particular attention to how the data were collected. It is possible that the results are legitimate, but it warrants your attention. Second, once you are satisfied you are dealing with a legitimate source, you can use them as a resource to understand the results.

The Condition of the Data

You can employ a few tools, available in statistical packages and spreadsheet software, to evaluate the condition of the data. For example, determine the minimum and the maximum values of the variables. Knowing these values is a first step to examine data quality.

If you were interested in an age variable and the maximum value was 110, the high value would warrant further investigation. It is possible that someone could be that old, but there are few instances of it. Another example might be that you are examining administrative data for a program that only enrolls children up to age 12, and you find the maximum value for age is 14. Again, it would merit your attention. By subtracting the minimum value from the maximum value, you can calculate the range. This will tell you how far apart the two values are.

The frequency distribution is another useful tool. It will tell you how often each value occurs for a given variable. The following table is an example of a frequency distribution.

GENERAL HEALTH				
	Frequency	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Excellent	1,136	23.51	1,136	23.51
Very Good	1,613	33.37	2,749	56.88
Good	1,388	28.72	4,137	85.60
Fair	495	10.24	4,632	95.84
Poor	190	3.93	4,822	99.77
Don't Know/ Not Sure	10	0.21	4,832	99.98
Refused	1	0.02	4,833	100.00

The first column contains the various values of the variable *General Health*. The second column, **Frequency**, refers to the number of times *General Health* takes a particular value. In this table 1,136 people reported their general health as "Excellent". The figure in the third column, **Percent**, is the percentage that the

frequency number is of the total number. **Cumulative Frequency**, the fourth column, is a running total of the frequencies of each value. For example, the total frequency of "Excellent" and "Very Good" is 2,749, which is the value of the **Cumulative Frequency** for the "Very Good" row. The **Cumulative Percent** found in the fifth column operates just as the **Cumulative Frequency** does except it calculates the cumulative value for **Percent**.

The frequency distribution can be used to:

- Examine the data for out-of-range values.
- Compare the distribution of values to an expected distribution.
- See if the total number of responses (cumulative frequency) is the correct number.
- Group data. For example, the cumulative frequency column reveals the combination of the "Excellent" and "Very Good" categories is 56.88% of the total.

These tools for assessing data can provide a great deal of insight into the quality of the data.

The Type of Data

Understanding the type of data in which you are interested is important as well; it will impact the ways in which you can report your data. There are four broad types of data:

1. Nominal
   It may be easier to think of nominal data as name data. The categories for this type of data are names. Some examples are counties, religious affiliation, or program name. There is no order to them.

2. Ordinal
   Ordinal data have some kind of order to them. Scales such as "Excellent," "Very Good," "Good," and "Poor" are ordinal data.

3. Interval
   Interval data are numbers. Temperature is an often cited example of interval level data. Mathematical functions can be performed on this kind of data because the distance between one value and another is measurable.

4. Ratio
   Ratio level data are the same as interval level data except that the value of zero means the absence of the characteristic. For example, if the weight of something is zero pounds then it has no weight. Compare temperature; if the temperature is zero, there is still a temperature, but we call it zero degrees.

An analytic plan is simply an outline of how you plan to analyze your data, or, if you are working with a data analyst, it is an outline of what kinds of data you will ask them to provide. At this point it, will be necessary to refine your question a little bit further to reflect how the data are collected. If the data are yours and you understand the way the data are collected, this will not be too difficult. However, it will be more difficult if you are unfamiliar with the data. If this is the case, work with the data analyst responsible for the data to refine your question to a point that it can be measured by that particular data system. This process is referred to as creating an "operational definition."

Once you have tailored your question(s), you can plan your analysis. This involves two steps:

1. Select your variables.
2. Select your methods.

Select your variables

This step includes not only the main variable you wish to analyze, but also those you want to use to examine your main variable. For example, if you are examining health care coverage, you may want to look at how health care coverage varies by income level, by geographic location, by sex, or by age. In order to visualize this, you can construct a table with the variables in it, such as the one below.

	Have Health Care Coverage	Do Not Have Health Care Coverage
State Total		
Sex		
Male		
Female		
Income		
<$15,000		
$15,000 - $50,000		
$50,000+		
Public Health District		
Public Health District 1		
Public Health District 2		
Public Health District 3		
Public Health District 4		

This table illustrates that the variables of interest are health care coverage by sex, income, and public health district.

Select Your Methods

Methods that are commonly used to summarize data are:

- Count.
- Rates (including percentages).
- Mean.
- Median.
- Mode.
- Frequency Distribution.

Count . The count is simply the number of things. It could be the number of people without health insurance, the number of clients that are served at a regional office in a given time period or even the number of clients served within a geographic area. When interpreting the count, it is important to keep in mind its size relative to the overall population. As the following table demonstrates, the count can increase, but if the population increases more, the count becomes a smaller proportion of the population. Therefore, care should be taken in the interpretation of counts. One could say that the count was increasing during the years 2000 – 2002, but would also have to say that its proportion of the population was decreasing over those same years.

	2000	2001	2002
Count	10	20	30
Population Size	100	250	400
Proportion of the Population	10%	8%	7.5%

It is for this reason that rates are calculated.

Rates. Rates are how often a particular event occurs in a given population. Percentages are probably the most common kinds of rates. The formula for calculating a percentage is:

$$\frac{\text{Count}}{\text{Population Size}} \times 100$$

A percentage is unique in that its base is 100. Rates with bases other than 100 are referred to as rates, but the base is always noted. It is common for very small numbers to be represented with rates per 1,000; 10,000; or even 100,000. The formula for a rate with a base of 100,000 would be:

$$\frac{\text{Count}}{\text{Population Size}} \times 100{,}000$$

Rates with bases of more than 100 are used when instances of a given event are infrequent, such as a rare cause of death. The value of rates is that they allow comparisons across different sizes of populations. This will often be the case when comparing data across years, geographic areas, and demographic groups. For example, you may be interested in how often binge drinking occurs among various age categories.

Age	**Binge Drinking**		
	Count	Population Size	Percent
18-24	75	300	25.0
25-34	160	800	20.0
35-44	135	900	15.0
45-54	100	1,000	10.0
55-64	30	600	5.0
65+	12	1,200	1.0

The first column of the table, **Age,** contains the age categories. The second column, **Count**, contains the number of those in the age group that are binge drinkers. The third column, **Population**, contains the total population of the age group. The fourth column, **Percent**, is the percentage of the age group that are binge drinkers. The percentages were calculated by dividing the number in the **Count** column (referred to as the "numerator") by the number in the **Population** column (referred to as the "denominator"). If you compare the numbers in the count column across the age groups, you may conclude that the "25-34" group has the biggest binge drinking problem. However, you will notice that the population numbers are quite different. Calculating the percentages creates a common base that allows a comparison across the age categories even though they have different populations. In this case, the "18-24" age category has the highest percentage of binge drinkers.

A useful tool when working with percentages is percentage-change. Percentage-change calculates how much change occurred in terms of a percentage. In order to calculate this figure, you need the old number and the new number. See the box below for the formula.

$$\frac{\text{New number} - \text{Old number}}{\text{Old number}} \times 100$$

For example, if the percentage of Idaho adults who have ever been told they have asthma changed from 10.8 percent to 11.7 percent, we could say that it increased by 8.3 percent.

The next three methods –the mean, median, and mode- are referred to as measures of "central tendency." The purpose of these measures is to indicate where the middle of the data is and to summarize data so that it can be compared. They can be interpreted to be "typical" or "expected" values for the variable.

The Mean. The mean is commonly known as the average. It is calculated by adding all the values for a particular variable and dividing by the number of observations. For example, suppose we want to look at program enrollment by region of enrollment.

	Total Program Enrollment
Region 1	200
Region 2	350
Region 3	400
Region 4	600
Region 5	450
Region 6	300
Region 7	150
Total Enrollment	2,450
Number of Regions	7
Formula for the mean	=2,450 ÷ 7
Mean enrollment	350

Program enrollment could be compared by comparing mean enrollment. This is illustrated in the following table.

	Total Program A Enrollment	Total Program B Enrollment
Region 1	200	300
Region 2	350	200
Region 3	400	450
Region 4	600	250
Region 5	450	200
Region 6	300	550
Region 7	150	410
Total Enrollment	2,450	2,360
Number of Regions	7	7
Formula for the mean	=2,450 ÷ 7	=2,360 ÷ 7
Mean (average) enrollment	350	337.1

The mean has an important limitation; it is influenced by extreme values. The following table demonstrates this.

	Total Program A Enrollment	Total Program B Enrollment
Region 1	100	100
Region 2	100	100
Region 3	100	100
Region 4	100	100
Region 5	100	100
Region 6	100	100
Region 7	1,000	100
Total Enrollment	1,600	700
Number of Regions	7	7
Formula for the mean	=1,600 ÷ 7	=700 ÷ 7
Mean (average) enrollment	228.6	100

Even though the enrollment numbers are all the same except for Region 7, the means are quite different. If all we knew about the program enrollment numbers were the means, then we might conclude that the numbers are very different when in fact only Region 7 is different. It is useful to know that there are extreme values in the data in order to

correctly interpret the means. You can identify the extreme values by calculating the minimum and maximum values. The following table illustrates.

	Program A Enrollment for Regions 1-7	Program B Enrollment for Regions 1-7
Mean (average) enrollment	228.6	100
Minimum Value	100	100
Maximum Value	1,000	100

Understanding the minimum and maximum values can help interpret the data. For example, because the minimum and maximum values for Program B are both 100, we know that all values for Program B are 100. The maximum value for Program A demonstrates that there is an extreme value which has the effect of increasing the mean value for that program.

The standard deviation is another tool for interpreting the mean. The mean doesn't reveal the degree of variation that exists in the data. The standard deviation measures the amount of variation in the data and how it varies around the mean. The formula for calculating the standard deviation is not presented here, but it can easily be calculated in a spreadsheet program like Excel®. For an example, we will examine the enrollment data again. This time we will analyze annual statewide enrollment.

Year	Total Program A Enrollment	Total Program B Enrollment
1999	2,000	4,000
2000	2,000	700
2001	2,000	400
2002	2,000	900
2003	2,000	4,000
Total Enrollment for 1999-2003	10,000	10,000
Number of Years	5	5
Formula for the mean	=10,000 ÷ 5	=10,000 ÷ 5
Mean (average) enrollment	5,000	5,000
Standard Deviation	0	1,640.7

The standard deviation is expressed in the same units as the data. The standard deviation is 0 for Program A because there is no variation in the annual values. It is much higher for Program B because of how varied the annual data is. For these data, it is easy to see the variation, but real data are rarely so uniform and there is usually so much data that the data cannot be evaluated merely by looking at it.

The Median. The median is also a tool used to find a "middle" value of a distribution of data. The median value is identified in the following way:

$$\frac{N+1}{2}$$

You then count the resulting number of places up or down in your data to find the median value. N represents the total count. Intuitively, we place data in order by category such as year or age category, but proper calculation of the median requires that the data values be in ascending or descending order. The following table illustrates.

Year	Clients Served by Program A
1999	8,500
2000	9,000
2001	10,000
2002	7,000
2003	8,000
Median value	10,000

In this case, the formula identifies 10,000 as the median value. It is easy to see that 10,000 is the maximum value, not the median value. To correct this, the data values (not to be mistaken with the data categories which are found in the "Year" column) must be placed in ascending or descending order.

Year	Clients Served by Program A
2002	7,000
2003	8,000
1999	8,500
2000	9,000
2001	10,000
Median value	8,500

You will note that identifying the median is not impacted by the values themselves. This means that the median is not subject to extreme values in the distribution. In this way, it corrects for the tendency of the mean to be influenced by extreme values. When there is an even number of observations, the median will fall between the two values. If that is the case, the average of the two values is the median value.

The Mode. The mode is the value that occurs most frequently. The following table presents the ages of clients served in Program A.

Client Ages Served by Program A	
27	35
35	46
25	32
56	41
62	35
35	62

In this example, the mode is 35 because it is the most frequently occurring value. It is also possible to have more than one mode.

Client Ages Served by Program A	
27	56
35	46
25	32
56	41
62	35
35	56

In this table the modes are 35 and 56.

The Frequency Distribution. The frequency distribution reports the relative frequency of each category. It can also show percent, rates, the cumulative frequency as well as the cumulative percent (see also page 8).

GENERAL HEALTH				
	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Excellent	1,136	23.51	1,136	23.51
Very Good	1,613	33.37	2,749	56.88
Good	1,388	28.72	4,137	85.60
Fair	495	10.24	4,632	95.84
Poor	190	3.93	4,822	99.77
Don't Know/ Not Sure	10	0.21	4,832	99.98
Refused	1	0.02	4,833	100.00

As the data are analyzed, exercise caution when dealing with small counts or small populations. Results from small numbers are more subject to random variation. Additionally, depending on the type of data, reporting data from small numbers, especially in a small geographic area, can result in a breach of confidentiality. Data can be grouped to larger geographic areas, larger demographic categories, or across years to gain larger numbers.

The count, percentage, rate, mean, median, mode, and frequency distribution are among the most common statistical measures. Which of these methods you select for your analysis is determined by what you want to know. It will also be determined by what kind of data you have; each type of data has methods that are appropriate for it. The table on the next page outlines the appropriate methods for each type of data.

Type of Data	Appropriate Methods
Nominal	Count Percentages Mode Frequency Distribution
Ordinal	Count Percentages Median Mode Frequency Distribution
Interval	Count Percentages (of values or collapsed categories) Mode(of values or collapsed categories) Frequency (of values or collapsed categories) Range Mean Standard Deviation
Ratio	Count Percentages (of values or collapsed categories) Mode(of values or collapsed categories) Frequency (of values or collapsed categories) Range Mean Standard Deviation

Once the analytic plan has been completed, the next step is to obtain the results. If you are working with an analyst, share your analytic plan with them. They will probably have additional questions and may be able to improve your plan once they understand what kinds of information you are looking for. If you are working with an analyst, keep in mind that the time it takes to analyze the data will vary with things like:

- The analyst's current work load.
- How familiar the analyst is with the data.
- How much quality checking of the data needs to be done.
- How much manipulating of the data needs to be done.
- The extent of the analytic plan.

The amount of output that an analyst returns to you is rarely a good reflection of the work it takes to get the results. The results are usually the tip of the iceberg.

If you are analyzing the data yourself, you may not have the analytic software tools that a statistical analyst might have. If not, many types of analysis can be done with spreadsheet software such as Excel®. There are some powerful tools in Excel® that are relatively easy to use.

AutoSum 

The AutoSum button can be found on the toolbar just below and to the right of Help. It looks like the symbol above. The Greek "E" is the mathematical symbol for sum. If you select a cell below a range of values and select the AutoSum button, it will create a total in that cell for the range of values. If you select the down arrow on the right hand side of the button, you will be given the options of Sum, Average, Count, Max (maximum), Min (minimum), and More functions.

Functions 

Excel® offers a wide selection of functions. Functions are mathematical operations, many of which are statistical functions. Just above the lettered column references in a worksheet is a formula bar (place your arrow over it and a label will pop up). Just to the left of the formula bar is an "f" and an "x" that look like the symbol above. If you select this button (or if you select the "More functions" item on the drop-down box of the AutoSum button) a pop-up box labeled "Insert Function" will pop up. It has a search feature and also a drop-down box that will let you select functions by category. The functions have good explanatory material with them.

Cells

The cells can also be programmed with mathematical operations by typing the formulae directly into the cells. Other useful tools include: "sort," "filter," "pivot tables," and "tables" in the "Data" menu. In the "Tools" menu try "Goal Seek," "Solver," and the "Data Analysis" tool pack (these last two have to be added by using

the Add-in function under the same menu). These and other tools will provide you with a wide range of analytic options.

Once the data have been analyzed, they should be summarized in some fashion. This section presents some general guidelines on presenting the data through technical writing, tables, graphs, and maps.

Technical Writing

As you begin summarizing your results, how you report them should be consistent with the question you asked at the start of the process. The purpose for which the original question was developed should guide you in how you present your results. The results can be a full-blown report, or they can be a summary report of several pages, an informal report for internal use, a single page fact sheet, or perhaps part of an email. Each type of reporting will have different requirements as to the degree of detail you include, but you should consider the following list of items as you assemble your reporting tool.

The text. As you write text that describes the data, be certain that you do not assume anything about the data that you do not know. It is easy to suggest causality between two variables or social reasons that the data may show this or that, but if the data do not explicitly demonstrate what you are trying to say, do not write it that way. It is important to the integrity of the process that unwarranted assumptions aren't made; it can damage your credibility if someone challenges assertions you have made about the data and you cannot back them up. Be sure that your final conclusions are consistent with your theoretical framework. Your framework places your data in context and will make your conclusions stronger.

Other elements you should consider as you report your results are methods, limitations, and sources. Each of these contribute to the transparency of your analytic process. They will allow others to evaluate your process and serve as a record for yourself and others that may use the data. If you have done a good job, others will emulate your methods.

Methods. The methods section will describe how you analyzed your data. This discussion can include the tools you used (including software), what analytic methods you used, and how you manipulated the data.

Limitations. Report any limitations that your data may have. This could result from the population from which the data were collected (for example, data collected only from adult males may not be generalizable to adult females or children), how the data were collected (for example, a telephone survey doesn't collect data from those without phones), the methods used to analyze the data, or the methods used to report the data. Much of this information can come from the assessment of the data.

Sources. Include the source(s) from which you obtained your data. You may also consider listing some of the more useful results of your literature review to aid others if they are researching the same topic.

Tables

A table is another useful tool for presenting data. Effective practices for including tables in documents are outlined here. The suggestions are meant as guidelines although a particular situation may require some deviation from them. Many of the suggestions in this section come from the *Gregg Reference Manual, Eighth Edition.*

Table Placement.

- The table should be placed on the same page that the subject is introduced.
- Avoid dividing a table between two pages.
- Center the table horizontally.
- Indent ½ of an inch, if possible.
- Insert 1 to 3 blank lines above and below the table.

If you have many tables, you can place them in the back as an appendix. If a table is not located on the same page as its reference, provide a cross reference in parenthesis.

Table Titles. Give each table a descriptive title. Keep title style consistent throughout the reporting document. The title should include enough information that will enable the reader to understand its contents independent of the surrounding text. This can include identification of the group(s), the variables, the date or time frame represented by the data, and the type of data (such as "percentages"). If you have many tables, you can number them for easy reference.

The Table.

- Provide a heading for each column in the table.
- Use singular forms of words for the column or row headings.
- In order to limit the size of columns, it is appropriate to use abbreviations and symbols.
- Text in column headers should be single spaced and centered.
- If the table is too large to place horizontally, place it on its own page vertically.

Table Text.

- The text within a table can be single spaced or double spaced, but this should be consistent throughout the document.
- If text takes up more than one line, indent the second line.
- Align text on the left and indent subentries.

Table Numbers.

- Align whole numbers (numbers without a decimal point) at the right.
- Align numbers with decimals by the decimal point.

- The column heading should indicate if your numbers are rates (including percentages).
- Totals can be very useful for the reader to interpret the data.

Table Notes.

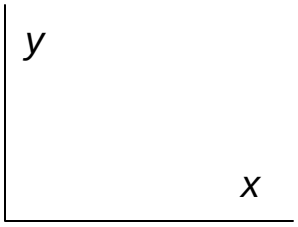
- Place notes at the bottom of the table.
- Notes can include explanatory information on missing items, definitions of terms, explanation of symbols, extreme data, or any other elements that might inform the reader.

Grouped data should be mutually exclusive, meaning that elements of the groups do not overlap. For example, if you were comparing two age groups – "18-64" and "65+", you should make certain that the two groups don't have common members. This could be the case if your two groups were "18-65" and "65+". In this example, it isn't clear which category is counting the age 65 population or if it is being counted twice. Everything should be consistent throughout the document and should facilitate the use of the data. Enough information should be presented in the table elements for a reader to understand it contents without referring to the text around it.

Graphs

Graphs (also called charts or figures) are useful for visually highlighting trends, differences, patterns, comparisons, and changes in data. Although technology has made the creation of graphs much simpler, it has also resulted in the over-engineering of graphs. This section is intended to explain the function of several common graphs and provide some guidance on their construction.

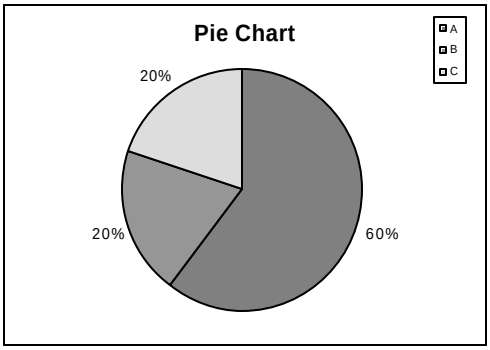
The Basics



Graphs are commonly constructed with two intersecting lines (the axes). The horizontal axis is called the x-axis, and the vertical axis is called the y-axis. Often categories such as income, place, sex, etc. are represented on the x-axis and the values for the categories are placed along the y-axis.

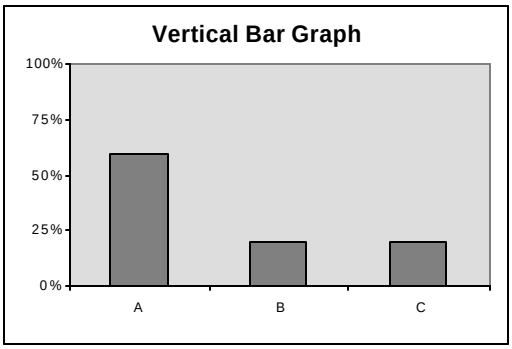
Commonly Used Graph Types

Pie Charts



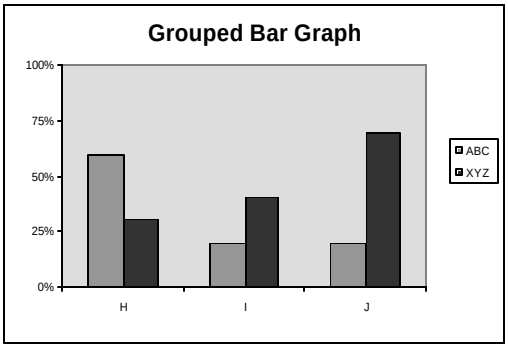
Pie charts are graphs that are pie-shaped – that is they are circular. The areas that represent the data are shaped like pieces of pie. Pie charts are used to show elements of a whole. For example, a pie chart could be used to show elements of a budget.

Vertical Bar Graphs



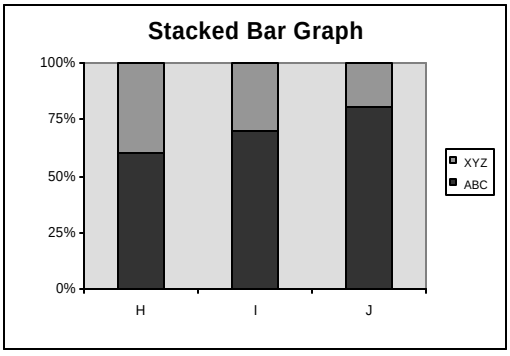
The regular vertical bar graph can be used for a variety of purposes including to show change over time and comparing a number of elements.

Grouped Bar Graphs



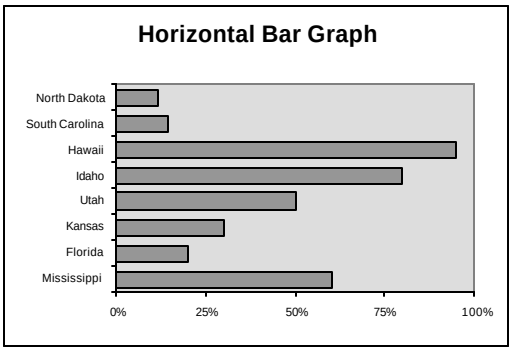
Grouped bar graphs are particularly useful for making comparisons among up to three groups. More groups than three can make your graph look cluttered.

Stacked Bar Graphs



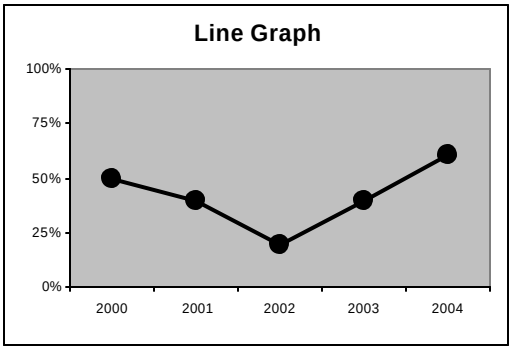
Stacked bar graphs enable the reader to see elements of a total and how they compare across categories such as time or age. As with all graphs, caution should be exercised with the number of elements that are graphed because more than a few will become difficult to read.

Horizontal Bar Graphs



Horizontal bar graphs are useful when there are too many categories to represent on the x-axis due to space limitations or when the category names are too lengthy.

Line Graphs



Line graphs effectively illustrate data over time. The lines readily show the continuity of time-related data. If more than one series of data is shown and some cross, use different patterns to distinguish the trends.

Graph Elements

Modern software packages allow for significant customization. These tools provide an opportunity to show data accurately and to good effect, but they can also result in cluttered graphs. The following guidelines, organized by graphical element, will help keep your graphs organized and clear. These are intended as guidelines only. Sometimes variation from these guidelines may be needed as the data and purpose dictate.

Titles. Titles should be concise. They should include identification of the group(s) described, the variables, the date or time frame represented by the data, and the type of data (such as "percentages"). The titles should be easy to read and understand, and they should be written horizontally.

Scales. The scale is the increment of the data shown along the y-axis. It should usually begin at zero. The maximum value should allow some white space on top. If you are presenting more than one graph on the same page or you intend for a set of graphs to be compared, they should have the same scale.

Patterns and Shading. Many different graphical elements can be altered with variations of patterns and shading. In general, patterns should be easy

on the eye; some available patterns look more like optical illusions and will be distracting. Patterns and shading should be easily distinguishable from each other. If line graphs cross each other, use different line patterns.

Axis Labels. Appropriately label the axes to inform the reader about the categories, scale, and number types (percentages, etc.).

Symbols. Symbols can be used to designate a data point on a line graph and they can be used in place of regular bars in bar graphs. As a rule, symbols should be kept simple and not distract from the data.

Color. Color is easily over-used. Use it sparingly; one or two colors can be used quite effectively.

All of these graph types and the options discussed here can be found in common spreadsheet software packages such as Excel®. In Excel®, find the symbol that looks like a graph (see example to the right) and click on it. You will find an easy to use interface that will provide you with many graphing options. This same symbol can be found in PowerPoint®. In Word®, select the "Insert" menu, then select "Object". Among the object types you can select to create graphs are Microsoft Excel Chart and Microsoft Graph Chart. Both can create graphs.

Remember to keep your graphs simple. The reason for the graph is to illustrate something about the data. Don't use graphical elements that will distract from that purpose. Sometimes several charts will serve better than one cluttered chart. Choose the type of graph and graph elements that will best suit your data and your purpose for presenting the data. Your graph should accurately reflect the data. You should not alter graphical elements to exaggerate or distort the data. Just as with tables, your categories should be mutually exclusive. Add appropriate titles and source material. Readers of the graph should be able to understand the graph without having to refer to the text before or after the graph.

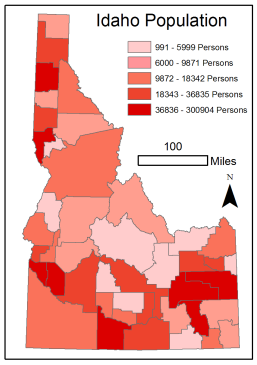
This checklist, found in the book *Graphing Statistics and Data*, will help you evaluate your work.

- Is the graph easy to read?
- Can the graph be misinterpreted?
- Does the graph have a good size and shape?
- Is the graph in the right place?
- Does the graph benefit from being in color?
- Have you tried the graph out on anybody?

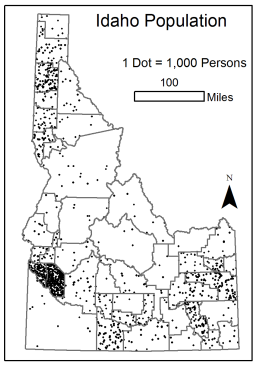
Maps

Maps are a powerful way to represent your data. Like tables and graphs, they can be used to simplify data and to reveal trends, differences, patterns, comparisons, and changes in data. Geographic Information Systems (GIS) Analysts produce maps by using commercially available software. This section discusses some of the more common types of maps and their uses.

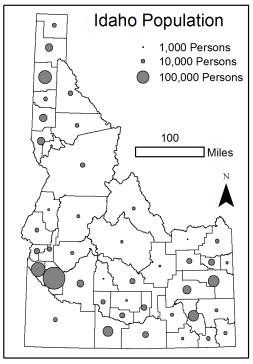
Choropleth maps. Choropleth maps depict geographic areas shaded according to data intensity. Data are assigned to mutually exclusive groups (also referred to as classes), and each group assigned a color shade. For example, assume a data set with values between 1 and 100. Possible groups are 1-25, 26-50, 51-75, and 76-100. Colors used to map data can consist of one primary color with progressively lighter or darker shades. This is called a color ramp. A primary color of red would have class colors between bright red and light pink (see example). Color ramps between three colors are also possible. A ramp from yellow to green to blue is an example; where yellow can be a negative value, green zero, and blue a positive value. Care must be taken when creating the classes, so as to depict the data as accurately as possible on the map.



Dot density maps. Dot density maps are randomly placed dots within a geographic area. Each dot represents a constant number of occurrences. Every dot on the map is the same size. For instance, a population map of a state by county might assign each dot to equal 100 persons. If a county has 10,000 residents it will be displayed containing 100 dots randomly distributed across the area of the county. These maps give the visual effect of density. It is important to be familiar with the data before making a dot density map. Knowing which areas should give the appearance of high density, and which areas will have low density is important to gauge the success of the map. A good dot density map should show dots coalescing in areas with high values. Notice in the example that Canyon and Ada counties appear almost solid. Dots should appear sparsely distributed in areas with low values. Dots that are too small will not show density. Dots that are too large will coalesce in areas that do not have large data values.



Proportional symbol maps. Proportional symbol maps use varying symbol size to represent the intensity of a value. The larger a proportional symbol is, the more intense the value. The cartographer (map maker) uses two variables to create proportional symbols: minimum and maximum symbol size. The smallest symbol size represents the smallest possible data value; the largest symbol size represents the largest data value. The rest of the symbol sizes are estimated between the minimum and maximum. Using proportional symbols is a good technique when mapping more than one variable. One variable can be displayed using symbol size while another variable is mapped using a colors. Point,



line, and area can be mapped using proportional symbols. Areas and points are symbolized with points, while linear features use line thickness.

The finished map should be easy to interpret for your target audience and should inform the reader about the data. A map should contain standard cartographic elements including a north arrow, scale, and a legend. A north arrow is used to orient the map reader. A scale bar allows the reader to comprehend distances depicted on the map. Legends inform the reader on how to interpret symbols on the map. The reader will not be able to interpret the content of the map if the legend is confusing or poorly designed. Just as with other methods of data display, the elements of the map should serve to communicate the data rather than confuse it.

Conclusions

This guide has presented the process of finding answers as a linear process; in practice however, you will find it is more of an iterative process. As you get more answers, you will often have new questions or the new information will cause you to reframe your original question.

One of most important keys to a successful analytic process is to maintain its integrity from beginning to end. This includes making the process transparent, selecting appropriate methods, making appropriate conclusions, and appropriately representing the data. Any method used to represent the data should be simple, clear, and honest in its presentation. As you are required to turn to data to support your program or public policy, these tools can be used to comprehend data and increase the quality of decision making.

Appendix A:
Data Sources

These data sources are a sampling of available data on the internet. There are many others; many government agencies have some kind of data posted on their websites.

State Data Sources

http://www.idcancer.org/
The Cancer Data Registry of Idaho produces reports on cancer within Idaho.

http://www.jobservice.ws/
The Idaho Department of Commerce and Labor website offers labor and workforce reports, and unemployment rates are available by county

http://www.idoc.state.id.us/data/index.html
The Idaho Department of Commerce and Labor website offers census, economic, community, international and tourism data.

http://www.corr.state.id.us/
The Idaho Department of Correction presents facts and figures on its inmate population.

http://www.sde.state.id.us/admin/statistics/
The State Department of Education has education statistics on Idaho's children.

http://www.sde.state.id.us/fedpro/hiv.asp
This **Idaho Department of Education** link provides results for the Youth Risk Behavior Survey, a health survey of young adults in high school.

http://www.deq.state.id.us/
The Idaho Department of Environmental Quality provides air and water quality as well as waste disposal information for the state of Idaho. This website includes a geographic information systems component.

http://www.healthandwelfare.idaho.gov/portal/alias__Rainbow/lang__en-US/tabID__3457/DesktopDefault.aspx
The Idaho Department of Health and Welfare has on-line health data publications that can be downloaded.

http://www.healthandwelfare.idaho.gov/portal/alias__Rainbow/lang__en-US/tabID__3424/DesktopDefault.aspx
This link has a listing of **Idaho Department of Health and Welfare** publications.

http://www.idahokidscount.org
Kids Count provides data on the economic, health, and education status of children.

http://www.isp.state.id.us/citizen/crime_stats.html
The Idaho State Police provide Idaho crime statistics through this website.

National Data Sources

http://www.ojp.usdoj.gov/bjs/welcome.html
The Bureau of Justice Statistics presents data on crime, victims, prosecution, the federal justice system, criminal offenders and other topics.

http://www.bls.gov/
The Bureau of Labor Statistics website presents data on inflation, consumer spending, wages, earnings, benefits, productivity, and unemployment.

http://www.cdc.gov/node.do/id/0900f3ec8000ec28
The Centers for Disease Control and Prevention compiles statistical information about the health of the nation. It collects data from birth and death records, medical records, interview surveys, and through direct physical exams and laboratory testing.

http://www.cms.hhs.gov/researchers/
The Centers for Medicare & Medicaid Services (CMS) is a Federal agency within the U.S. Department of Health and Human Services. Programs for which CMS is responsible include Medicare, Medicaid, and the State Children's Health Insurance Program (SCHIP).

http://www.fbi.gov/ucr/ucr.htm#cius
Federal Bureau of Investigation presents results from The Uniform Crime Reporting (UCR) Program.

http://www.fedstats.gov/
The Fedstats website is a clearinghouse of federal statistics websites.

http://www.aecf.org/kidscount/data.htm,
Kids Count provides data on the economic, health and education status of children.

http://www.usda.gov/nass/
The National Agricultural Statistics Service (U.S. Department of Agriculture) has a variety of types of agriculture data.

http://www.nimh.nih.gov/healthinformation/statisticsmenu.cfm
The National Institute of Mental Health offers statistics on various mental health disorders for the nation as a whole.

http://oas.samhsa.gov/states.htm
The Substance Abuse and Mental Health Services Administration provides state level data and trends on substance abuse, mental health, and access to treatment.

http://www.census.gov/
The U.S. Census Bureau presents detailed population, demographic and housing data, maps and summaries from United States Census 2000.

Blankenship, A. G., George Edward Breen and Alan Dutka. 1998. State of the Art Marketing Research. 2d ed. Chicago, Illinois: NTC Business Books.

Dunn, William N. 1994. Public Policy Analysis: An Introduction. 2d ed. Englewood Cliffs, New Jersey: Prentice Hall.

Groebner, David F. and Patrick W. Shannon. 1981. Business Statistics: A Decision Making Approach. Columbus, Ohio: Charles E. Merrill Publishing Company.

Handler, Arden, Deborah Rosenberg, Colleen Monahan, and Joan Kennelly, eds. 1998. Analytic Methods in Maternal and Child Health. Maternal and Child Health Bureau, Health Resources and Services Administration, Department of Health and Human Services.

Hedrick, Terry E., Leonard Bickman, and Debra J. Rog. 1993. Applied Research Design: A Practical Guide. Newbury Park, California: Sage Publications, Inc.

Leaverton, Paul E. 1991. A Review of Biostatistics: A Program for Self Instruction. Boston, Massachusetts: Little, Brown and Company.

Milstein, Robert L. and Scott F. Wetterhall. 1999. Framework for Program Evaluation in Public Health. Morbidity and Mortality Weekly Report. September 17: vol 48 (RR11). pp. 1-40.

Mitchell, Andy. 1999. The ESRI Guide to GIS Analysis. Redlands, California: Environmental Systems Research Institute.

Patton, Carl V. and David S. Sawicki. 1993. Basic Methods of Policy Analysis and Planning. Englewood Cliffs, New Jersey: Prentice Hall.

Sabin, William A. 1996. The Gregg Reference Manual. 8th ed. Westerville, Ohio: Glencoe/McGraw Hill.

Schlotzhauer, Sandra D. and Ramon C. Littell. 1997. SAS® System for Elementary Statistical Analysis. 2d ed. Cary, North Carolina: SAS Institute Inc.

Shively, W. Phillips. 1990. The Craft of Political Research. 3d ed. Englewood Cliffs, New Jersey: Prentice Hall.

Sirkin, Mark A. 1995. Statistics for the Social Sciences. Thousand Oaks, California: Sage Publications, Inc.

Slocum, Terry A. 1999. Thematic Cartography and Visualization. Upper Saddle River, New Jersey: Prentice Hall.

Thompson, Nancy J. and Helen O. McClintock. 1998. Demonstrating Your Program's Worth: A Primer on Evaluation for Programs To Prevent Unintentional Injury. Atlanta,

Georgia: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control.

Tufte, Edward R. 2001. The Visual Display of Quantitative Information. 2d ed. Cheshire, Connecticut: Graphics Press.

Wallgren, Anders, Britt Wallgren, Rolf Persson, Ulf Jorner, and Jon-Aage Haaland. 1996. Graphing Statistics and Data: Creating Better Charts. Newbury Park, California: Sage Publications, Inc.

Wheeler, Gloria. Basic Applied Statistics for Public Management. Self Published Manuscript.

White, Louise G. 1990. Political Analysis: Technique and Practice. 2d ed. Pacific Grove, California: Brooks/ Cole Publishing Company.